2025-03-28

causalReg: Integrating Regression-based Causal Inference Tools for Empirical Research

Dylan Junluo CHEN

Department of Information Systems, College of Business, City University of Hong Kong

Signatories

Project team

- **Dylan Junluo CHEN** (Primary Investigator) PhD Student, Department of Information Systems College of Business, City University of Hong Kong Research interests: open source community, data breaches, empirical methods Homepage: https://chen-junluo.com/
- **Bofei SHE** PhD Student, College of Computing National University of Singapore Expertise: R programming, econometric methods, statistical analysis

Contributors

- **Prof. Ben LIU** Department of Information Systems College of Business, City University of Hong Kong Contribution: Econometric methodology guidance and validation
- **Prof. Xin LI** Department of Information Systems College of Business, City University of Hong Kong Contribution: Machine learning methodology consultation

Consulted

- **Fellow PhD students** at the Department of Information Systems, City University of Hong Kong who use R as their primary analysis tool for empirical research and have expressed interest in contributing to the open source community.
- **Members of the China-R community** have been consulted regarding the needs of Chinese empirical researchers and the potential impact of this package.

The Problem

In quantitative empirical research within China's academic community, Stata has long dominated as the primary tool for econometric analysis, particularly for causal inference techniques such as two-way fixed effects models and various difference-in-differences (DID) implementations (Angrist and Pischke 2009). This creates several critical challenges:

Limited Methodological Access

While Stata offers established packages for traditional econometric approaches, newer methodologies often appear first in R. This creates a significant gap for Chinese research students who rely primarily on Stata. For example, recent developments in:

1. **Omitted variable bias sensitivity analysis** (Oster 2019) - While implemented in Stata, current implementations don't fully support multiple fixed effects structures common in modern empirical work.
2. **Heterogeneous treatment effect estimation** - The SortedEffects package in R (Chernozhukov, Fernández-Val, and Luo 2018) exhibits critical bugs with continuous variables that prevent proper analysis, yet remains unmaintained.
3. **Optimal policy learning** - Tools for determining which samples should receive which treatments to maximize effectiveness remain fragmented and inaccessible to non-specialists.

Technical Barriers to Adoption

For many Chinese research students, transitioning from Stata to R presents significant challenges:

1. Language barriers limit access to English-language tutorials and documentation
2. Lack of integrated teaching materials specifically targeting econometric applications
3. Absence of comprehensive Chinese-language support for implementing causal inference methods in R

Previous Attempts and Limitations

Several R packages individually address specific causal inference techniques, but they lack:

1. **Integration** - No single package consolidates regression-based causal inference methods
2. **Accessibility** - Limited bilingual documentation targeting Chinese users
3. **Maintenance** - Critical bugs in packages like SortedEffects remain unaddressed
4. **Pedagogy** - Missing connection to the applied econometrics workflow familiar to Stata users

Why This Problem Matters

This situation creates a significant opportunity cost for China's empirical research community:

1. **Research quality** - Limited access to cutting-edge methods affects the rigor of empirical work
2. **Methodological innovation** - Restricted adoption of newer approaches hampers methodological advancement
3. **Research efficiency** - Working around tool limitations consumes valuable research time

4. **Global collaboration -** The divergence in tooling creates barriers to international research partnerships

The R ecosystem offers substantial advantages for causal inference with its open-source nature, reproducibility, visualization capabilities, and rapid integration of methodological innovations. Creating a bridge between this ecosystem and China's research community would significantly advance empirical research capabilities.

The proposal

Overview

The causalReg package will provide an integrated toolkit for regression-based causal inference methods in R, specifically designed to facilitate adoption by China's empirical research community. By addressing the methodological gaps identified in the problem statement, causalReg will:

1. **Integrate critical causal inference techniques** into a cohesive framework built on established R packages like lfe (Gaure 2013)
2. **Provide bilingual documentation and tutorials** targeting Chinese research students
3. **Establish ongoing maintenance** for key capabilities including SortedEffects
4. **Create a bridge** between Stata-centric workflows and R's powerful ecosystem

This project directly aligns with the R Consortium's mission by expanding R's reach in a significant academic community, ensuring package sustainability, and promoting reproducible research practices.

Detail

The causalReg package will be developed in three progressive phases, each targeting a key component of the regression-based causal inference workflow:

Phase 1: Omitted Variable Bias Sensitivity Analysis

Building on the methodology developed by Oster (2019), we will implement a robust solution for sensitivity analysis that:

- Extends existing implementations to support multi-way fixed effects through integration with the lfe package
- Provides intuitive visualization of sensitivity bounds
- Includes comprehensive documentation with applied examples from economics and information systems

The implementation will handle both continuous and binary outcomes, with full support for the standard econometric workflow including clustered standard errors and various fixed effects specifications.

Phase 2: Treatment Effect Heterogeneity Analysis

We will revive and enhance the SortedEffects package (Chernozhukov, Fernández-Val, and Luo 2018) to:

- Fix critical bugs affecting continuous variable analysis
- Improve integration with standard regression workflows
- Add robust visualization capabilities for heterogeneous effects
- Enhance documentation with practical examples

This module will allow researchers to move beyond average treatment effects to understand how causal impacts vary across the population - a critical capability for policy analysis.

Phase 3: Policy Learning Implementation

Building on recent methodological developments in machine learning for causal inference (Athey and Wager 2021), we will:

- Implement methods for optimal treatment targeting
- Provide tools to estimate heterogeneous treatment effects with machine learning
- Create accessible interfaces for policy learning without requiring deep ML expertise
- Develop step-by-step tutorials for applied policy analysis

Minimum Viable Product

The MVP for causalReg will include:

1. A functioning R package with robust implementations of Oster sensitivity analysis and SortedEffects
2. Comprehensive bilingual documentation including:
    - Function references
    - Methodological vignettes
    - Applied tutorials with real-world datasets
3. Integration with existing R econometric workflows
4. Initial educational materials for Chinese users

Architecture

The package will be built with a modular design that:

1. Leverages established packages (lfe, sandwich, ggplot2) for core functionality
2. Provides consistent syntax across different causal inference methods
3. Includes flexible visualization components
4. Offers both high-level interface functions and access to underlying components

This architecture ensures maintainability while allowing advanced users to customize analyses as needed.

Assumptions

Our approach is built on several key assumptions:

1. The primary barrier to R adoption among Chinese research students is the lack of integrated, well-documented causal inference tools
2. The proposed methodologies represent the most pressing needs for this community
3. Bilingual materials will significantly reduce adoption barriers
4. The integration of these methods will provide sufficient value to motivate users to transition from Stata

Project plan

Start-up phase (Month 1)

The initial phase will focus on establishing project infrastructure and planning:

- Set up GitHub repository with proper contribution guidelines
- Establish code style, documentation standards, and review processes
- Create continuous integration pipeline for testing
- Finalize detailed technical specifications for each module
- Develop initial project website structure
- Select MIT license for maximum accessibility and reuse

Technical delivery

Phase 1: Omitted Variable Bias Sensitivity Analysis (Month 2-3)

Month 2: Development - Implement core sensitivity analysis functionality based on Oster methodology - Integrate with lfe package for fixed effects support - Develop visualization components - Create test suite with comprehensive coverage

Month 3: Documentation and Release - Complete English and Chinese function documentation - Develop vignette explaining methodology and application - Prepare example dataset and use cases - Release initial version (0.1.0) to GitHub - Submit to CRAN

Phase 2: Treatment Effect Heterogeneity Analysis (Month 4-5)

Month 4: Development - Fork and repair SortedEffects package - Fix continuous variable handling bug - Enhance integration with standard regression workflows - Improve visualization capabilities

Month 5: Integration and Documentation - Integrate with causalReg package - Complete bilingual documentation - Develop tutorial vignette with applied examples - Release updated version (0.2.0)

Phase 3: Policy Learning Implementation (Month 6-7)

Month 6: Development - Implement core policy learning algorithms - Create interface for optimal treatment targeting - Develop visualization for policy insights - Build test suite for functionality

Month 7: Final Integration and Release - Integrate policy learning module with package - Complete comprehensive documentation - Develop advanced tutorial vignette - Release complete package version (1.0.0)

Other aspects

Community Building and Promotion (Ongoing)

- **Month 3**: Initial announcement blog post and social media (WeChat) campaign
- **Month 5**: Webinar demonstrating first two modules (recorded in Chinese and English)
- **Month 7**: Comprehensive video tutorial series
- **Month 8**: Workshop materials for university training sessions

Reporting and Communication

- Monthly progress reports to ISC
- Quarterly blog posts on R Consortium website
- Final project report and case study

Requirements

People

Core Team

The project will be led by Dylan Junluo CHEN (PhD student, City University of Hong Kong), who brings: - Experience in empirical research methods and causal inference - R package development knowledge - Background in both information systems and economics research

The core team also includes: - **Bofei SHE** (PhD student, NUS College of Computing): Providing programming expertise and econometric knowledge - **Prof. Ben LIU** (City University of Hong Kong): Offering econometric guidance and methodology oversight - **Prof. Xin LI** (City University of Hong Kong): Providing machine learning expertise, particularly for the policy learning module

Additional support will come from PhD colleagues at CityU with experience in R programming and empirical research who have expressed interest in contributing to open-source projects.

Roles and Responsibilities

- **Dylan Junluo CHEN**: Project management, core development, documentation, community engagement
- **Bofei SHE**: Development support, code review, testing
- **Prof. Ben LIU**: Econometric methodology consultation
- **Prof. Xin LI**: Machine learning methodology consultation

Processes

The project will implement the following processes:

1. **Development Workflow**:
   - GitHub-based development with branch protection
   - Pull request review requirements (minimum 1 reviewer)
   - Continuous integration testing
   - Documentation updates required for all new features
2. **Communication**:
   - Bi-weekly team meetings
   - Monthly progress updates to ISC
   - Public development roadmap on GitHub
   - Issue tracker for bug reports and feature requests
3. **Community Engagement**:
   - WeChat community group for Chinese users
   - Regular office hours for user questions
   - Translation workflow for maintaining bilingual documentation
4. **Quality Assurance**:
   - Comprehensive test suite covering all functions
   - Real-world use case validation
   - User feedback incorporation process

Tools & Tech

The project will utilize:

1. **Development Tools**:
   - R (>= 4.0.0) as primary programming language
   - RStudio for development environment
   - roxygen2 for documentation
   - testthat for unit testing
   - GitHub Actions for continuous integration
   - pkgdown for website generation
2. **Dependencies**:
   - lfe: Core engine for fixed effects regression
   - ggplot2: Visualization
   - sandwich: Robust standard errors
   - glmnet: For certain machine learning components
   - dplyr and other tidyverse tools for data manipulation
3. **Infrastructure**:
   - GitHub for code hosting
   - CRAN for distribution
   - GitHub Pages for documentation website

Funding

The requested funding of $1,000 will be allocated as follows:

1. **Educational Materials ($300)**:
   - Video tutorial production
   - Example dataset preparation
   - Translation services for documentation
2. **Community Building ($300)**:
   - Virtual workshop hosting
   - WeChat official account maintenance
   - User community support
3. **Contingency ($400)**:
   - Reserved for unexpected requirements

Summary

This project requires a focused development team with expertise in both econometrics and R programming, supported by academic advisors with domain knowledge. The technical requirements are modest, primarily relying on existing R infrastructure and packages. The funding requested will primarily support dedicated development time to ensure timely completion of all components and the creation of high-quality educational materials necessary for adoption by the target user community.

Success

Definition of done

The causalReg project will be considered complete when:

1. A fully functional R package is published on CRAN containing all three modules:
   - Omitted variable bias sensitivity analysis
   - Treatment effect heterogeneity analysis
   - Policy learning implementation
2. Comprehensive bilingual documentation is available including:
   - Complete function reference (English and Chinese)
   - Methodological vignettes explaining the underlying theory
   - Applied tutorials with real-world examples
   - Video tutorials covering key workflows
3. A sustainable community structure is established:
   - Active WeChat group for Chinese users
   - Responsive GitHub issue process
   - Clear contribution guidelines for future development
4. Educational materials are available:
   - Workshop materials for university courses
   - Self-paced learning resources
   - Example datasets with complete workflows

Measuring success

We will track project success through several quantifiable metrics:

Short-term metrics (1 year): - GitHub repository: 100+ stars - CRAN downloads: 1,000+ in first year - WeChat articles: 5+ with 1,000+ views each - Video tutorials: 5+ with 500+ views each - Community size: 100+ members in WeChat group - Issue resolution: 90% of reported bugs resolved within 2 weeks

Medium-term metrics (2 years): - Academic citations: 5+ papers using the package - Educational adoption: 3+ university courses using the materials - Package extensions: 2+ external contributions merged - CRAN downloads: 5,000+ cumulative

Qualitative success indicators: - Positive feedback from Chinese research students - Reduction in reported barriers to R adoption - Examples of research using methods previously inaccessible

Future work

This project lays the foundation for several future extensions:

1. **Additional methodologies**:
   - Synthetic control methods
   - Instrumental variable approaches
   - Regression discontinuity designs

- Machine learning-based causal inference techniques

2. **Integration with other ecosystems**:
   - Interoperability with Python tools like DoWhy
   - Connections to Bayesian frameworks
   - Integration with tidymodels ecosystem
3. **Educational expansion**:
   - Full online course development
   - Textbook companion materials
   - Integration with university curricula
4. **Community development**:
   - Annual contributor workshops
   - Regional user groups in major Chinese cities
   - Regular webinar series featuring applied case studies

Key risks

People

Risk: Limited availability of core team members due to academic commitments **Mitigation**: Clear time allocation agreements, modular development allowing parallel work, comprehensive documentation to facilitate onboarding additional contributors

Risk: Difficulty finding contributors with both econometric and R expertise **Mitigation**: Focused recruitment efforts in academic communities, development of contributor guides that separate technical and methodological tasks

Processes

Risk: Challenges in maintaining bilingual documentation synchronization **Mitigation**: Establish clear translation workflow, prioritize function documentation and core vignettes, leverage community for translation assistance

Risk: Scope creep as additional methodologies are requested **Mitigation**: Maintain strict prioritization process, establish clear criteria for inclusion in core package versus extensions

Tooling & Technology

Risk: Compatibility issues with future R versions **Mitigation**: Implement comprehensive continuous integration testing, monitor R-devel changes, establish deprecation policies for API changes

Risk: Dependency on lfe package maintenance **Mitigation**: Document key dependencies, establish contingency for potential forking if necessary, maintain communication with upstream package maintainers

Costs

Risk: Underestimation of development time requirements **Mitigation**: Phase-based approach allows for reassessment at milestone points, prioritization of core functionality first

Risk: Unexpected costs for community building and outreach **Mitigation**: Leverage existing academic networks, utilize free platforms where possible, allocate contingency funds

Angrist, Joshua D, and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton university press.

Athey, Susan, and Stefan Wager. 2021. "Policy Learning with Observational Data." *Econometrica* 89 (1): 133–61.

Chernozhukov, Victor, Iván Fernández-Val, and Ye Luo. 2018. "Sorted Effects Method: Discovering Heterogeneous Effects Beyond Their Averages." *Econometrica* 86 (6): 1911–38.

Gaure, Simen. 2013. "Lfe: Linear Group Fixed Effects."

Oster, Emily. 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics* 37 (2): 187–204.